JOURNAL OF ENGINEERING SCIENCES
ЖУРНАЛ ІНЖЕНЕРНИХ НАУК
ЖУРНАЛ ИНЖЕНЕРНЫХ НАУК
Web site: http://jes.sumdu.edu.ua
Volume 4, Issue 1 (2017)

УДК 681.327.12

Детектування мовної активності в автоматизованій системі розпізнавання мовця критичного застосування

М. М. Биков[1)], В. В. Ковтун[1)], О. О. Максимов[1)]

[1)] *Вінницький національний технічний університет, Хмельницьке шосе, 95, 21021, м. Вінниця, Україна*

Article info:
Paper received: April 18, 2017
The final version of the paper received: May 25, 2017
Paper accepted online: May 29, 2017

Correspondent Author's Address:
vntu0113055@gmail.com

У статті автори розробили метод детектування мовної активності для автоматизованої системи розпізнавання мовців критичного застосування із вейвлет-параметризацією мовного сигналу та класифікацією на інтервали «мова»/«пауза» з використанням згортальної нейромережі. Запропонований авторами метод вейвлет-параметризації дозволяє обрати оптимальні параметри вейвлет-перетворення відповідно до заданої користувачем похибки подання мовного сигналу. Також метод дозволяє здійснювати оцінювання втрат інформації залежно від вибраних параметрів неперервного вейвлет-перетворення (НВП), що дозволило зменшити кількість обчислюваних коефіцієнтів НВП мовного сигналу на порядок із допустимим ступенем спотворення локального спектра НВП. Також запропоновано алгоритм детектування мовної активності із згортальним нейромережевим класифікатором, який показує високу якість сегментації мовних сигналів на інтервали «мова»/«пауза» та є стійким до присутності у мовному сигналі вузькосмугового шуму і техногенних шумів за рахунок властивостей згортальної нейромережі.

Ключові слова: автоматизована система розпізнавання мовців критичного застосування, детектування мовної активності, вейвлет-перетворення, згортальна нейромережа.

1. ВСТУП

Детектор мовної активності (voice activity detector) – це алгоритм, призначений для розпізнавання інтервалів активної мови і пауз. Такі алгоритми активно використовуються в системах стільникового зв'язку, де на часових інтервалах пауз, передавання сигналу, як правило, не відбувається, а передається лише загальна інформація про фон. У сфері сучасних прикладних інформаційних систем найбільш часто із завданням детектування мовної активності зіштовхуються у галузі стільникової телекомунікації, тому природно, що саме у цій галузі створено найбільше алгоритмів детектування мовної активності, які пройшли процедуру стандартизації із залученням відповідних профільних організацій. Зокрема, Міжнародний комітет ITU-T випустив детектори G.729 Annex B (G.729B) [1] і G.723.1 Annex A (G.723.1A) для можливості переривчастого передавання мовного сигналу (Discontinuous transmission – DTX), Європейський інститут стандартизації ETSI рекомендував алгоритми детектування мовної активності GSM-FR, -HR і -EFR для європейських систем цифрового стільникового зв'язку [1]. Згодом ETSI представив адаптивні детектори мовної активності AMR1 і AMR2 [1], для використання у мережах третього покоління стандарту UMTS. Північноамериканські організації зі стандартизації TIA&EIA представили алгоритми IS-96 і IS-127 [1]. Найчастіше у запропонованих алгоритмах як інформаційні ознаки для детектування інтервалів пауз використовуються енергії сигналів у виділених частотних смугах і спектральна форма сигналу, проте розглянемо принципи функціонування згаданих алгоритмів детальніше.

В алгоритмі детектування мовної активності G.729B запис мовного сигналу розбивається на інтервали тривалістю 10 мс. Віднесення сигналу на поточному інтервалі до одного із класів «мова»/«пауза» відбувається на підставі аналізу значень чотирьох характеристичних параметрів – різниці енергій усього частотного діапазону, різниці енергій у діапазоні низьких частот, спотворення спектра у десяти частотних смугах і різниці кількостей переходів амплітудою сигналу нульового. Мовний сигнал на поточному часовому інтервалі належить до класу «мова» якщо більшість із характеристичних параметрів перевищують задані порогові рівні. Алгоритм детектування мовної активності G.723.1A функціонує аналогічно вищеописаному і відрізняється від G.729B лише тривалістю часових інтервалів аналізу мовного сигналу, яка для алгоритму G.723.1A дорівнює 30 мс.

В алгоритмах детектування мовної активності GSM-FR/HR/EFR, запропонованих ETSI, як характеристична ознака при класифікації мовного сигналу використовується залишкова енергія, значення якої під час прийняття рішення порівнюється з адаптивним порогом. Прогнозована залишкова енергія обчислюється як абсолютне значення різниці дійсної енергії сигналу та прогнозованого її значення. Для

одержання прогнозованих значень енергії сигналу використовується автокореляційна функція.

В алгоритмі детектування мовної активності AMR1 утворюється смуговий фільтр, що розділяє вхідний сигнал на дев'ять нерівномірних частотних смуг, де нижні частотні смуги перекривають менший частотний діапазон, порівняно зі смугами вищих частот. Згідно з алгоритмом для кожної частотної смуги обчислюється енергія сигналу та оцінюється відношення сигнал/шум, для чого за допомогою авторегресійної моделі першого порядку розраховується енергія фонових шумів. Рішення щодо класу аналізованого інтервалу вхідного сигналу приймається на підставі узагальнення результатів порівняння суми оцінок відношень сигнал/шум для всіх частотних смуг з адаптивним пороговим значенням. На відміну від алгоритму AMR1 в алгоритмі AMR2 застосовується швидке перетворення Фур'є вхідного сигналу, який розбивають на 16 частотних смуг із нерівномірним покриттям частотного діапазону – ширина смуги збільшується із зростанням частоти. Використовуючи спектрограми вхідного сигналу і фонового шуму, обчислюється відношення сигнал/шум для кожної частотної смуги. Для опису енергії фонового шуму застосовується авторегресійна модель першого порядку. Для уникнення надмірної чутливості до нестаціонарних фонових шумівв алгоритмі проводиться оцінювання дисперсії миттєвих значень міжінтервальних значень відношень сигнал/шум. Рішення щодо класу інтервалу вхідного сигналу приймається на підставі відношень пікових значень відношень сигнал/шум до середнього.

Асоціації TIA&EIA представили алгоритми детектування мовної активності IS-96 і IS-127, де вхідний сигнал розбивається на дві частотні смуги з подальшим розрахунком енергії на часових інтервалах аналізу, значення яких використовується для обчислення авторегресійної функції прогнозування значень енергії. Смуга нижніх частот використовується для оцінювання фонового шуму, тоді як смуга верхніх частот використовується для детектування мовного сигналу. Рішення щодо класу інтервалу вхідного сигналу приймається на підставі порівняння відношень сигнал/шум із значеннями адаптивних порогів, що залежать від рівня фонового шуму і відношення сигнал/шум попереднього інтервалу.

Результати тестування вищеописаних алгоритмів свідчать, що алгоритми детектування мовної активності ідентифікують інтервали мови і пауз з різною ефективністю. Так, алгоритм G.729B найточніше серед інших ідентифікує паузи, однак найгірше ідентифікує інтервали мови, для алгоритму IS-127 картина зворотна. Алгоритми GSM-EFR, AMR1 і AMR2 показують близькі результати класифікації інтервалів мови, а результати детектування інтервалів пауз значно відрізняються. Алгоритм GSM-EFR найефективніше з-поміж інших сегментує мовний сигнал при співвідношенні сигнал/шум більше ніж 15 дБ, але із зменшенням шуму показує істотно гірші результати. Найбільш стійким до рівня шуму у сигналі виявися алгоритм AMR2, якість сегментації для якого залишалася на середньому рівні для всіх досліджуваних рівнів сигнал/шум.

У класичній теорії розпізнавання образів [2] також вирішується завдання детектування мовної активності, яку ще називають сегментацією мовного сигналу. Відомі алгоритми сегментують мовний сигнал на інтервали «мова»/«пауза» на підставі даних енергії сигналу і кількості переходів через нульовий рівень. Така параметризація мовного сигналу не враховує особливостей частотного спектра корисного сигналу і шуму, отже, не завжди дозволяє правильно класифікувати інтервали, особливо за наявності у сигналі вузькосмугового шуму або музичного фону. Загалом використання спектрального аналізу для детектування мовної активності не є оптимальним, оскільки одержання спектра Фур'є пов'язано з використанням усього інтервалу аналізованого сигналу, а короткочасний спектр Фур'є або має обмежену роздільну здатність у частотному просторі, або потребує використання завеликих інтервалів аналізу у часовому просторі.

Аналіз наведеної інформації показав, що жоден із досліджуваних алгоритмів не дозволяє здійснювати стабільну та якісну сегментацію мовного сигналу на класи «мова» і «пауза» одночасно, що зумовлює необхідність створення авторського методу сегментації для застосування в автоматизованих системах розпізнавання мовців.

Отже, враховуючи специфіку автоматизованих систем критичного застосування, необхідно створити метод оцінювання втрат інформації про індивідуальні особливості мовного сигналу під час виконання процедури детектування мовної активності із виділенням інтервалів «мова»/«пауза» у мовному сигналі. Для цього необхідно розробити метод параметризації мовного сигналу з можливістю оцінювання похибки його представлення, сформулювали алгоритм класифікації на його підставі та провести емпіричне дослідження одержаних теоретичних результатів.

2. ОСНОВНА ЧАСТИНА

2.1. Метод параметризації мовного сигналу для подальшого детектування мовної активності

Завдання автоматизованого детектування мовної активності передбачає параметризацію мовного сигналу для виявлення інтервалів «мова»/«пауза» і нормалізації тривалості звучання мовних сигналів. Таку операцію можна виконати, зокрема, застосувавши неперервне вейвлет-перетворення (НВП), що дозволить обрати бажану роздільну здатність відображення локальних спектрів мовного сигналу маніпулюючи параметрами материнського вейвлета. Необхідно відзначити, що одержані в результаті внаслідок вейвлет-перетворення значення зсуву і масштабу повинні максимально повно характеризувати мовний сигнал, проте їх кількість варто мінімізувати, формуючи лаконічний вектор характеристичних ознак і заощаджуючи обчислювальні ресурси.

У загальному випадку НВП сигнал $u(t)$ подається процедурою згортання [3–6]:

$$W_u(a,b)=\int_{-\infty}^{\infty}u(t)\cdot\psi\left(\frac{t-b}{a}\right)dt\,, \qquad (1)$$

де a – масштаб (безрозмірна величина, обернено пропорційна частоті); b – координати зсуву (у часовому просторі); $\psi(t,a,b)$ – двопараметрична вейвлет-функція (материнський вейвлет).

За материнський вейвлет у подальших дослідженнях автори обрали вейвлет Морле [4], оскільки його частотно-часові характеристики аналогічні характеристикам базилярної мембрани слухової системи людини. Для застосування вейвлету Морле необхідно задавати параметр масштабу σ, який визначає розмір вікна аналізу, і параметр домінантної частоти ξ, який дозволяє варіювати вибірковістю базису. Змінюючи значення цих параметрів, можна досягти бажаної ширини частотного і часового вікон (параметр σ) та високої точності апроксимації, використовуючи невелику кількість коефіцієнтів вейвлет-перетворення, внаслідок резонансу сигналу із вейвлетом (параметр ξ).

Вейвлет Морле за умови $\xi>4$ забезпечує збереження нульового середнього і згасання зі зростанням частоти спектральних складових сигналу (перетворення Фур'є, ПФ) материнського вейвлету. Його аналітичний вигляд такий:

$$\psi(t)=\frac{1}{\sqrt{\sigma\sqrt[4]{\pi}\sqrt{|a|}}}e^{j\xi}\cdot e^{-\frac{t^2}{2\sigma^2}}\,. \qquad (2)$$

Тоді НВП (1) для вейвлету Морле з урахуванням (2) набирає вигляду [5, 6]:

$$W_u(a,b)=\frac{1}{\sqrt{\sigma\sqrt[4]{\pi}\sqrt{|a|}}}\int_{-\infty}^{\infty}u(t)\cdot exp\left(-\frac{(t-b)^2}{2\sigma^2a^2}-j\xi\frac{t-b}{a}\right)dt\,. \qquad (3)$$

Для материнського вейвлету Морле існує вираз, що встановлює зв'язок між масштабом і частотою: $a=\frac{\omega_н}{\omega}=\frac{f_н}{f}$, де $\omega_н$ і $f_н$ – нормовані колова та лінійна частоти відповідно, які для вейвлету Морле розраховують за формулами $\omega_н=\frac{\xi+\sqrt{\xi^2+2\sigma^{-2}}}{2}$ і $f_н=\frac{\xi+\sqrt{\xi^2+2\sigma^{-2}}}{4\pi}$. У результаті поєднаємо параметри σ і ξ із a співвідношеннями $\sigma a=\sigma\frac{\omega_н}{\omega}=\frac{\xi\sigma+\sqrt{(\xi\sigma)^2+2}}{2\omega}=\frac{\beta+\sqrt{\beta^2+2}}{2\omega}$, $\frac{\xi}{a}=\xi\frac{\omega_н}{\omega}=\frac{2\omega}{1+\sqrt{1+2\beta^{-2}}}$, де $\beta=\xi\sigma$ – параметр, що характеризує НВП. При значенні параметра $\beta>2\sqrt{5}\approx4.472$ виконується умова $\xi^2\gg2\sigma^{-2}$, що дозволяє застосовувати наближені вирази $\sigma a\approx\beta/\omega$, $\xi/a\approx\omega$. І, в свою чергу, спрощує (3) до такого вигляду:

$$W_u(a,b)=\frac{1}{\sqrt[4]{\pi}}\sqrt{\frac{\omega}{\beta}}\int_{-\infty}^{\infty}u(t)\cdot e^{-\frac{(t-b)\cdot\omega^2}{2\beta^2}}\cdot e^{-j(t-b)\omega}dt\,, \qquad (4)$$

де функція $e^{-\frac{(t-b)\cdot\omega^2}{2\beta^2}}=h(t-b,\omega)$ описує часове вікно змінної ширини і залежить від частоти.

Для оброблення сигналів за умови $\xi>4.5$ значення параметра β за замовчуванням беруть таким, що дорівнює 5, однак залежно від необхідної роздільної здатності за частотою або в часі параметр β може набувати інших значень.

Для оброблення мовного сигналу, заданого множиною відліків $u_i=u(t_i)=u(i\Delta)$, $i\in0,\ldots,N-1$, де N – кількість відліків, $\Delta=1/f_d$ і f_d – крок і частота дискретизації відповідно, безпосереднє обчислення інтеграла НВП $W_u(a,b)$ чисельними методами, крім проблеми збіжності для малих значень a, потребує значних обчислювальних ресурсів. Для швидкого обчислення НВП мовного сигналу можна скористатися алгоритмом, який ґрунтується на рівності Парсеваля [7]:

$$W_u(a,b)=\frac{\sqrt{a_m}}{2\pi}\int_{-\infty}^{\infty}U(\omega)\cdot\Psi_\omega(k.m)\cdot e^{j\omega b}d\omega\,, \qquad (5)$$

де $U(\omega)$ і $\Psi(\omega)$ – перетворення Фур'є $u(t)$ і $\psi(t)$ відповідно. У дискретній формі вираз (5) для значень a_m і b_n набирає вигляду

$$W_u(m,n)=\frac{\sqrt{a_m}}{2\pi}\sum_{k=0}^{N-1}C_k\cdot\Psi_\omega(k.m)\cdot e^{j\frac{2\pi nk}{N}}\,, \qquad (6)$$

де C_k - коефіцієнти дискретного перетворення Фур'є (ДПФ) для відліків мовного сигналу u_i, які обчислюються за алгоритмом швидкого перетворення Фур'є (ШПФ). Відліки зсуву зазвичай задають лінійною залежністю $b_n=b_{min}+\Delta bn$, $n=0,1,\ldots,N_b-1$, $\Delta b=\frac{b_{max}-b_{min}}{N_b-1}$.

Вираз (6) визначає алгоритм розрахунку НВП, в якому зворотне ДПФ обчислюється за допомогою процедури БПФ. Перетворення Фур'є для вейвлету Морле (2) має вигляд $\Psi(\omega a)=\sqrt{2\pi}\sigma e^{-\frac{\sigma^2(\xi-\omega a)^2}{2}}$, що дозволяє одержати функцію $\Psi_\omega(k,m)=\sigma\sqrt{2\pi}e^{-\frac{\sigma^2}{2}\left(\xi-\frac{2\pi ka_m}{N\Delta}\right)^2}$.

Значення масштабувального параметра вейвлету a пропонується оптимізувати для заощадження обчислювальних ресурсів при здійсненні процедури параметризації мовного сигналу, для чого, враховуючи його природу, пропонується використати математичний апарат нелінійної апроксимації.

Визначимо бажану відносну похибку кроку частотної сітки для параметризації мовного сигналу за умови зміни масштабувального параметра a як $\varepsilon=\frac{\Delta f_{cep}}{f_0}=\frac{0.5(\Delta f_+ + \Delta f_-)}{f_0}=\frac{0.5((f_0-f_{0+})-(f_{0-}-f_0))}{f_0}$, де f_{0-}, f_0, f_{0+} – значення частоти спектральної складової мовного сигналу при a_{m-1}, a_m, a_{m+1} значеннях масштабувального параметра відповідно, $\Delta f_+, \Delta f_-, \Delta f_{cep}$ –

абсолютні помилки при збільшенні і зменшенні масштабуючого параметра і середнє їх значення відповідно. Відносну похибку параметризації мовного сигналу з урахуванням зміни масштабувального параметра для узагальненої апроксимувальної функції визначимо виразом $\psi = \frac{a_m(a_{m+1} - a_{m-1})}{2a_{m+1}a_{m-1}}$. Тоді критерій для визначення припустимої похибки в частотному просторі при зміні масштабувального параметра має вигляд $\psi \le \theta$.

Сформулюємо умови, які повинна забезпечити функція, що апроксимує зміни масштабувального параметра:

1. Монотонне зростання функції – $a_{m+1} > a_m$.

2. Монотонне зростання масштабуючого параметра – $a_{m+1} - a_m > a_m - a_{m-1}$ або $0.5(a_{m+1} + a_{m-1}) > a_m$.

3. Значення відносної похибки кроку частотної сітки не повинно перевищувати задану помилку – $\frac{a_m(a_{m+1} - a_{m-1})}{2a_{m+1}a_{m-1}} \le \theta$.

Пересвідчимося, що показникова функція, яка апроксимує зміни масштабувального параметра, $a_m = a_0 C^{dm}$, $m = 0, 1, \ldots, M$ задовольняє сформульовані вище умови при значенні ступеня C: $1 < C \le e = 2.71828 \forall d > 0$, де d - константа в показнику ступеня. У загальному вигляді показникові апроксимувальні функції можна подати у вигляді $a_m = a_0 2^{Em} = a_0 e^{\varepsilon m}$, де константи в показниках пов'язані відношенням $E = \varepsilon / ln2$, $a_0 = \Delta / \Delta_t$ – мінімальне значення масштабу вейвлету, де Δ_t – ефективний розмір материнського вейвлета у часовому просторі, який для вейвлету Морле дорівнює $\Delta_t = \sigma / \sqrt{2}$. Загалом $a_0 = \sqrt{2} / \sigma f_d$.

Номер найбільшого відліку масштабного параметра розраховується за формулою

$$M = \left\lceil E^{-1} log_2\left(\frac{a_M}{a_0}\right) \right\rceil = \left\lceil \varepsilon^{-1} ln\left(\frac{a_M}{a_0}\right) \right\rceil = \left\lceil \varepsilon^{-1} ln\left(\varphi \frac{f_d}{f_{min}}\right) \right\rceil, \qquad (7)$$

де $\lceil \ \rceil$ – закруглення до більшого цілого числа; f_{min} – мінімальна основна частота в спектрі мовного сигналу; $\varphi = f_u \Delta_t = \sigma f_u 2^{-0.5} = \frac{\sigma\xi + \sqrt{\sigma^2\xi^2 + 2}}{4\sqrt{2}\pi} = \frac{\beta + \sqrt{\beta^2 + 2}}{4\sqrt{2}\pi}$ – константа для вейвлету Морле.

Значення M-го відліку масштабного параметра розраховується за формулою

$$a_M = \frac{\sqrt{2}\varphi}{\sigma f_{min}} = a_0 \varphi \frac{f_d}{f_{min}}. \qquad (8)$$

Оцінити перевагу від застосування показникової функції, що апроксимує зміни масштабувального параметра порівняно з лінійною, можна відношенням $V = N_a^{лін} / M$, де кількість відліків масштабувального параметра при лінійній функції апроксимації $N_a^{лін} = \frac{a_{max} - a_{min}}{\Delta a} = \frac{a_M - a_0}{\Delta a}\Big|_{a_M >> a_0} \approx \frac{a_M}{\Delta a}$, де Δa – крок масштабного параметра для максимальної частоти при відносній похибці кроку частотної сітки ε; подамо як $\Delta a = a_1 - a_0 = \frac{\sqrt{2}}{\sigma(f_d - \varepsilon f_d)} - \frac{\sqrt{2}}{\sigma f_d} = \frac{\sqrt{2}}{\sigma f_d} \frac{\varepsilon}{1 - \varepsilon}\Big|_{\varepsilon << 1} \approx \frac{\sqrt{2}\varepsilon}{\sigma f_d}$. Тоді $N_a^{лін} = \left\lceil \frac{1}{\varepsilon} \varphi \frac{f_d}{f_{min}} \right\rceil$ і вираз для обчислення V такий: $V = \varphi \frac{f_d}{f_{min}} \Big/ ln\left(\varphi \frac{f_d}{f_{min}}\right)$.

Зменшити кількість коефіцієнтів НВП мовного сигналу можна, здійснюючи їх проріджування, тобто використовуючи для параметризації мовного сигналу лише ті коефіцієнти НВП, які йдуть через p відліків (з усього набору відліків сигналу $i \in 0, N-1$) Під час проріджування в p разів крок зсуву для обчислюваних коефіцієнтів НВП розраховуватимемо за формулою $\Delta b = p\Delta$, де p – коефіцієнт проріджування. Кількість відліків зсуву після процедури проріджування опишемо відношенням $N_b = \lceil N/p \rceil$. Введемо показник спотворення локального спектра НВП при проріджуванні в p разів Q_W, який описуватимемо відношенням

$$Q_W = \frac{\sum_{m=0}^{M} \sum_{i=0}^{N-1} \left(E^{вих}(m,i) - E^{пр}(m,i)\right)^2}{\sum_{m=0}^{M} \sum_{i=0}^{N-1} \left(E^{вих}(m,i)\right)^2} \cdot 100\ \% \qquad (9)$$

де $E^{вих}(m,i)$ і $E^{пр}(m,i)$ – відповідно вихідний і проріджений локальний спектр НВП у логарифмічному масштабі.

2.2. Емпіричні дослідження детектування мовної активності із застосуванням авторського методу параметризації мовного сигналу

Експериментальне дослідження ефективності запропонованого авторами алгоритму детектування мовної активності згортальною нейромережею [8] із вищеописаним методом параметризації мовного сигналу проводилося на підставі мовного матеріалу із безкоштовної бази еталонних записів NOIZEUS [9] – спеціалізованої бази даних Школи інжинірингу та комп'ютерних наук Еріка Джонсона при Університеті Техасу в Далласі, США, яка використовується для дослідження алгоритмів покращання звуку і складається з 30 речень англійської розмовної мови, вимовлених трьома чоловіками та трьома жінками (по 5 на кожного мовця, частота дискретизації записів становить 25 кГц, але для додавання шуму була зменшена до 8 кГц) та записів типових побутових і техногенних шумів.

Оцінювання втрати інформації з мовного сигналу унаслідок його проріджування обчислювалося за допомогою виразу (9) для мовного сигналу із частотою дискретизації $f_d = 8$ кГц та загальною тривалістю $N = 3000$ відліків. Одержані результати у вигляді залежності коефіцієнту проріджування p від кількості відліків зсуву після процедури проріджування N_b та показника спотворення локального спектра НВП Q_W подано на рисунку 1.

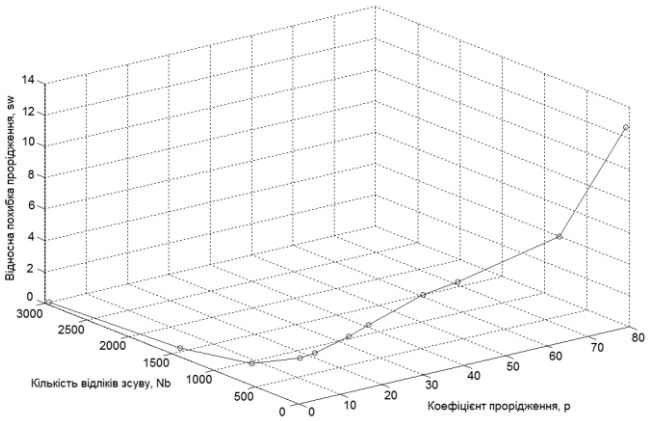

Рисунок 1 – Залежність втрати інформації з мовного сигналу від ступня його проріджання

Як бачимо з рис. 1, локальний спектр НВП мовного сигналу при значенні коефіцієнта $p \le 10$ зазнає незначних спотворень ($Q_w \le 2$), що дозволяє зменшити на порядок кількість обчислюваних коефіцієнтів НВП і заощадити апаратні ресурси системи.

Синтез алгоритму детектування мовної активності на підставі методу проріджання локальних спектрів НВП проведемо з урахуванням особливостей сприйняття звуку людиною, які описуються перцептуальною моделлю [10].

Модель розділяє спектр мовного сигналу на критичні частотні смуги. Кожна критична смуга частотного діапазону мовлення за компонентою шуму в ній сприймається як єдине ціле, тому для адекватного слухового сприйняття важливою є потужність шуму в смузі.

Відповідно до перцептуальної моделі частотний діапазон 0–23 500 Гц розбивається на 25 критичних смуг, ширина яких поступово зростає від 100 до 8 000 Гц [10]. Якщо діапазон мовного сигналу обмежений частотою 8 кГц, то можна або обмежитися 22 критичними смугами, або використати стандартні алгоритми підвищення частоти дискретизації, зокрема до 22 кГц, із незначним спотворенням вхідного сигналу. Враховуючи критичне застосування автоматизованої системи розпізнавання мовців, у якій використовуватиметься синтезований алгоритм детектування мовної активності, автори обрали перший варіант і використали для фільтрації мовного сигналу смуговий фільтр із 22 критичних смуг перцептуальної моделі.

Алгоритм детектування мовної активності для автоматизованої системи розпізнавання мовців критичного застосування складається із таких етапів:

1. Одержання сегмента мовного сигналу тривалістю 10 мс.
2. Проходження сегментом мовного сигналу смугового фільтра, створеного із урахуванням перцептуальної моделі.
3. Обрання параметрів НВП з урахуванням (9).
4. Одержання коефіцієнтів НВП для сигналів із кожної зі смуг фільтра та обчислення їх потужності.
5. Передавання візуального зображення потужності коефіцієнтів НВП на згортальну нейромережу для прийняття рішень щодо належності аналізованого сименсу до класів «мова»/«пауза».
6. Перехід до наступного сегмента.

Інфографіка описаного алгоритму із видаванням основних параметрів згортальної нейромережі наведена на рисунку 2 (кількість смуг фільтра – 10).

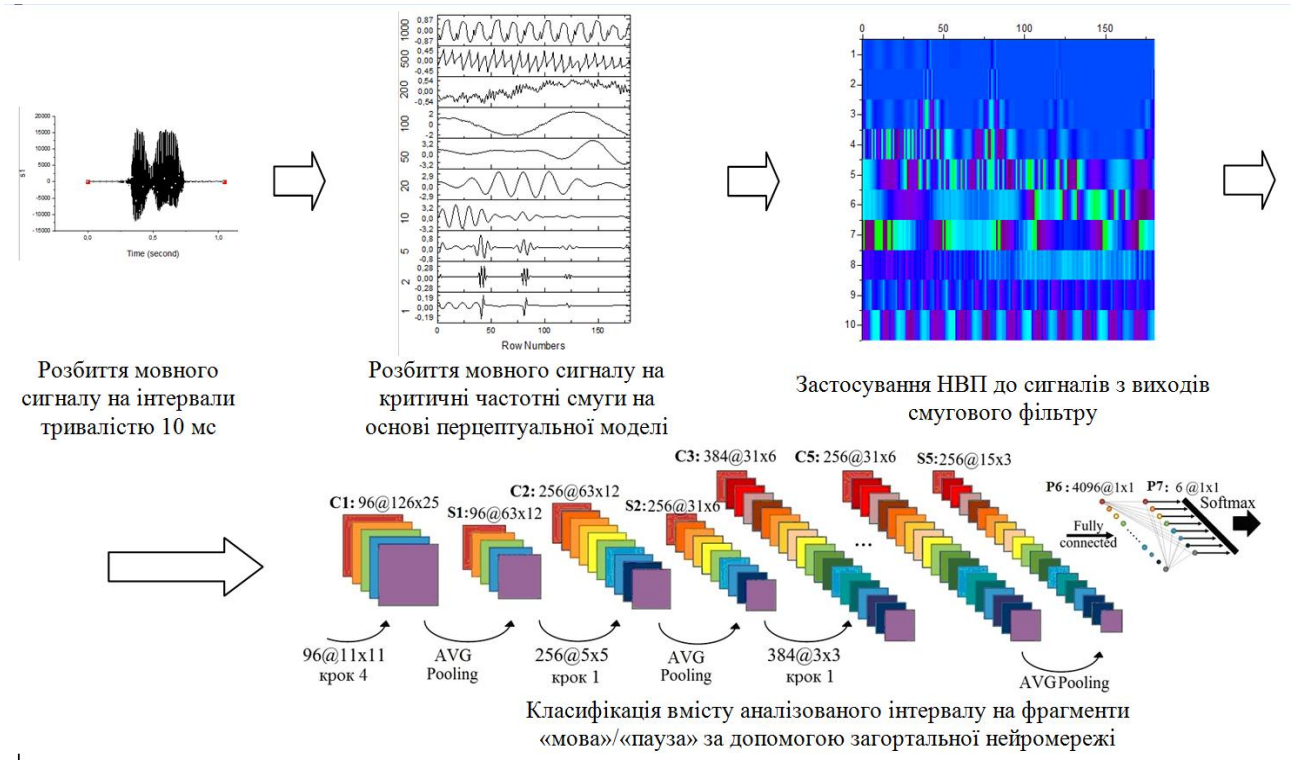

Рисунок 2 – Схема алгоритму сегментації мовного сигналу на інтервали «мова»/«пауза» на підставі локального спектра НВП

Результати тестування запропонованого алгоритму наведені на рисунку 3. Для тестування використовувався матеріал бази еталонних записів NOIZEUS без додавання до мовного сигналу шуму та із додаванням техногенних шумів до одержання співвідношення шум/сигнал 5, 10, 15 дБ відповідно. Якість алгоритму розраховувалася як відношення тривалості правильно розпізнаних сегментів до загальної тривалості звучання тестового мовного матеріалу.

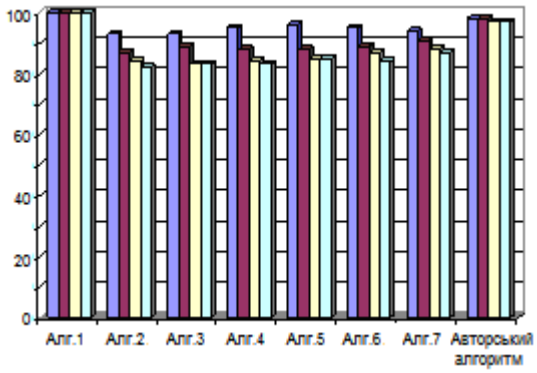

Рисунок 3 – Порівняння якості алгоритмів детектування мовної активності

Як бачимо із рисунка 3, авторський алгоритм показує найвищу якість сегментації мовного матеріалу на відрізки «мова»/«пауза», проте тривалість генерування висновку авторського алгоритму у 5–6 разів триваліша порівняно з конкурентами. Втім, роботу авторського алгоритму можна пришвидшити знизивши вимоги до якості параметризації мовного сигналу (збільшити інтервал проріджения p), але це призведе до погіршення якості сегментації.

3. ВИСНОВКИ

Автори дослідили питання детектування мовної активності в контексті використання в автоматизованій системі розпізнавання мовців критичного застосування. В роботі запропоновано метод параметризації мовного сигналу для виявлення інтервалів «мова»/«пауза» і нормалізації тривалості звучання мовних сигналів із динамічним вибором параметрів вейвлет-перетворення відповідно до інтенсивності локального спектра НВП. Також метод дозволяє оцінювати втрати інформації залежно від вибраних параметрів неперервного вейвлет-перетворення, що дозволило зменшити кількість обчислювальних коефіцієнтів НВП мовного сигналу на порядок при значенні показника спотворення локального спектра НВП менше ніж 2 %. Застосування вейвлет-перетворення для параметризації мовного сигналу дозволило з потрібною чутливістю здійснювати аналіз мовного сигналу у частотному та часовому діапазонах і з урахуванням запропонованого методу проріджения мовного сигналу виявляється більш обчислювально ефективним порівняно з традиційним аналізом Фур'є.

Створений на підставі запропонованого методу вейвлет-параметризації мовного сигналу алгоритм детектування мовної активності із згортальним нейромережевим класифікатором показує високу якість сегментації мовних сигналів на інтервали «мова»/«пауза» і є стійким до наявності у мовному сигналі вузькосмугового шуму та техногенних шумів за рахунок властивих згортальній нейромережі властивостей. До недоліків обраного способу класифікації можна віднести значну ресурсомісткість на етапі навчання класифікатора, яка, втім, компенсується запропонованим авторами методом проріджения локального спектра НВП.

Speech activity detection for the automated speaker recognition system of critical use

M. M. Bykov[1)], V. V. Kovtun[1)], O. O. Maksimov[1)]

[1)] *Vinnytsia National Technical University, 95 Khmelnytske Av., 21021, Vinnytsia, Ukraine*

In the article, the authors developed a method for detecting speech activity for an automated system for recognizing critical use of speeches with wavelet parameterization of speech signal and classification at intervals of "language"/"pause" using a curvilinear neural network. The method of wavelet-parametrization proposed by the authors allows choosing the optimal parameters of wavelet transformation in accordance with the user-specified error of presentation of speech signal. Also, the method allows estimating the loss of information depending on the selected parameters of continuous wavelet transformation (NPP), which allowed to reduce the number of scalable coefficients of the LVP of the speech signal in order of magnitude with the allowable degree of distortion of the local spectrum of the LVP. An algorithm for detecting speech activity with a curvilinear neural network classifier is also proposed, which shows the high quality of segmentation of speech signals at intervals "language" / "pause" and is resistant to the presence in the speech signal of narrowband noise and technogenic noise due to the inherent properties of the curvilinear neural network.

Keywords: automated speaker recognition system of critical use, speech activity detection, wavelet transformation, convolution neural network.

Детектирование речевой активности в автоматизированной системе распознавания диктора критического применения

Н. М. Быков[1)], В. В. Ковтун[1)], А. А. Максимов[1)]

[1)] *Винницкий национальный технический университет, Хмельницкое шоссе, 95, 21021, г. Винница, Украина*

В статье авторы разработали метод выявления речевой деятельности для автоматизированной системы распознавания критического использования языков с вейвлет-параметризацией речевого сигнала и классификации с интервалами «языка»/«паузы» с помощью криволинейной нейронной сети. Предложенный авторами метод вейвлет-параметризации позволяет выбирать оптимальные параметры вейвлет-преобразования в соответствии с заданной пользователю ошибки представления речевого сигнала. Также метод позволяет оценить потерю информации в зависимости от выбранных параметров непрерывного преобразования вейвлета (АЭС), что позволило уменьшить количество масштабируемых коэффициентов ЛВП речевого сигнала в порядке величины с допустимой степенью искажения локальный спектр LVP. Также предложен алгоритм обнаружения речевой активности с классифицируемой криволинейной нейронной сетью, показывает высокое качество сегментации речевых сигналов с интервалами «язык»/«пауза» и устойчива к присутствиям в речевом сигнале узкополосных шумов и техногенных шумов, за счет свойств, обладаемых криволинейной нейронной сетью.

Ключевые слова: автоматизированная система распознавания дикторов критического применения, детектирования речевой активности, вейвлет-преобразования, сверточная нейросеть.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Kondoz A. M. Digital Speech. Coding for Low Bit Rate Communication Systems / A. M. Kondoz. – John Wiley & Sons, Ltd., 2004. – 442 p.
2. Рабинер Л. Р. Цифровая обработка речевых сигналов / Л. Р. Рабинер, Р. В. Шафер. – Москва : Радио и связь, 1981. – 593 с.
3. Короновский А. А. Непрерывный вейвлет-анализ и его приложения / А. А. Коронский, А. А. Храмов. – Москва : Физматлит, 2003. – 176 с.
4. Горшков Ю. Г. Новые решения речевых технологий безопасности / Ю. Г. Горшков // Специальная техника – 2006. – № 4. – С. 1–13.
5. Huang X. Spoken language processing: a guide to theory, algorithm, and system development / X. Huang, A. Acero, H. Hon. – Prentice Hall PTR, 2001. – P. 936.
6. Добеши И. Десять лекций по вейвлетам / И. Добеши ; пер. с англ. – Ижевск : НИЦ «Регулярная и хаотическая динамика», 2001. – 464 с.
7. Бурнаев Е. В. Применение вейвлет преобразования для анализа сигналов / Е. В. Бурнаев. – Москва : МФТИ, 2007. – 138 с.
8. CS231n: Convolutional Neural Networks for Visual Recognition [Електронний ресурс]. – Режим доступу : http://cs231n.github.io/convolutional-networks.
9. NOIZEUS: Noisy speech corpus – Univ. Texas-Dallas [Електронний ресурс]. – Режим доступу : http://ecs.utdallas.edu/loizou/speech/noizeus.
10. Рабинер Л. Теория и применение цифровой обработки сигналов / Л. Рабинер, Б. Гоулд ; пер. с англ. – Москва : Мир, 1978. – 848 с.

REFERENCES

1. Kondoz, A. M. (2004). Digital Speech. Coding for Low Bit Rate Communication Systems. John Wiley & Sons, Ltd.
2. Rabiner, L. R., Shafer, R. V. (1981). Tsyfrovaia obrabotka rechevykh syhnalov [Digital processing of speech signals]. Moskow, Radio i sviaz [in Russian].
3. Koronovskiy, A. A., Khramov, A. A. (2003). Nepreryvnyi veivlet-analyz y eho prylozhenyia [Continuous wavelet analysis and its applications]. Moskow, Fizmatlit [in Russian].
4. Horshkov, U. H. (2006). Novye reshenyia rechevykh tekhnolohyi bezopasnosty [New solutions of speech safety technologies]. Spetsyal'naya tekhnika, 4, 1–13.
5. Huang, X., Acero, A., Hon, H. (2001). Spoken language processing: a guide to theory, algorithm, and system development. Prentice Hall PTR, 936.
6. Dobeshi, I. (2001). Desiat lektsyi po veivletam (perevedeno s anhlyiskoho) [Ten lectures on wavelets (translated from English)]. Izhevsk, NIC [in Russian].
7. Burnaev, E. V. (2007). Prymenenye veivlet preobrazovanyia dlia analyza syhnalov [Wavelet transformation for signal analysis]. Moskow, MFTI [in Russian].
8. CS231n: Convolutional Neural Networks for Visual Recognition. Retrieved from http://cs231n.github.io/convolutional-networks.
9. NOIZEUS: Noisy speech corpus - Univ. Texas-Dallas. Retrieved from http://ecs.utdallas.edu/loizou/speech/noizeus.
10. Rabiner, L., Hould, B. (1978). Teoriya i primenenie tsyfrovoi obrabotki signalov [Theory and application of digital signal processing]. Moscow, Mir [in Russian].